

Overview of OpenSRS

TOWARD AN OPEN IMPLEMENTATION OF ISO 11238

Tyler Peryea, Noel Southall, Trung Nguyen

February 4, 2013

NCATS

The proposed implementation is a cultivation of our efforts over the years to *develop, demonstrate, and disseminate* software for managing, mining, and visualizing chemical biology data.

Outline

- ISO 11238 and scope
- Reference substance database
- Software requirements
- Implementation roadmap
- Architecture overview
- Technology stack
- Status & milestones
- Discussion (and demo)

ISO 11238 and Scope

- Substance categories
 - » Chemical
 - » Protein
 - » Nucleic acid
 - » Polymer
 - » Structurally diverse
- Specified substances Groups 1–4
- Official names in multiple languages, jurisdictions, and domains
- Well-defined references and relationships between substances
- Unique identifiers

Reference Substance Database

- A reference substance database is distributed with each OpenSRS deployment
 - » Bootstrap from FDA's public SRS records
- A public accessible “master” copy of the database is housed at NCATS
 - » Data curation
 - » Conflict resolution
- Defined update schedule

Software Requirements

- Self-contained and modular
 - » Run entirely on a desktop or access remotely
- Well-defined data access application programming interface (API)
 - » Mobile or third-party clients
- Fine-grained security model
 - » Access control for every piece of information
- Audit trail of all data fields
- Web- and desktop-based clients
 - » Multiple platforms (e.g., Linux, Windows, Mac) for desktop client
- Basic support for text, structure, sequence searching
- Wizard interface to guide registration
- Support attachments (e.g., PDF's, MS spectra, images)
- Configurable “business rules” for standardizing structures

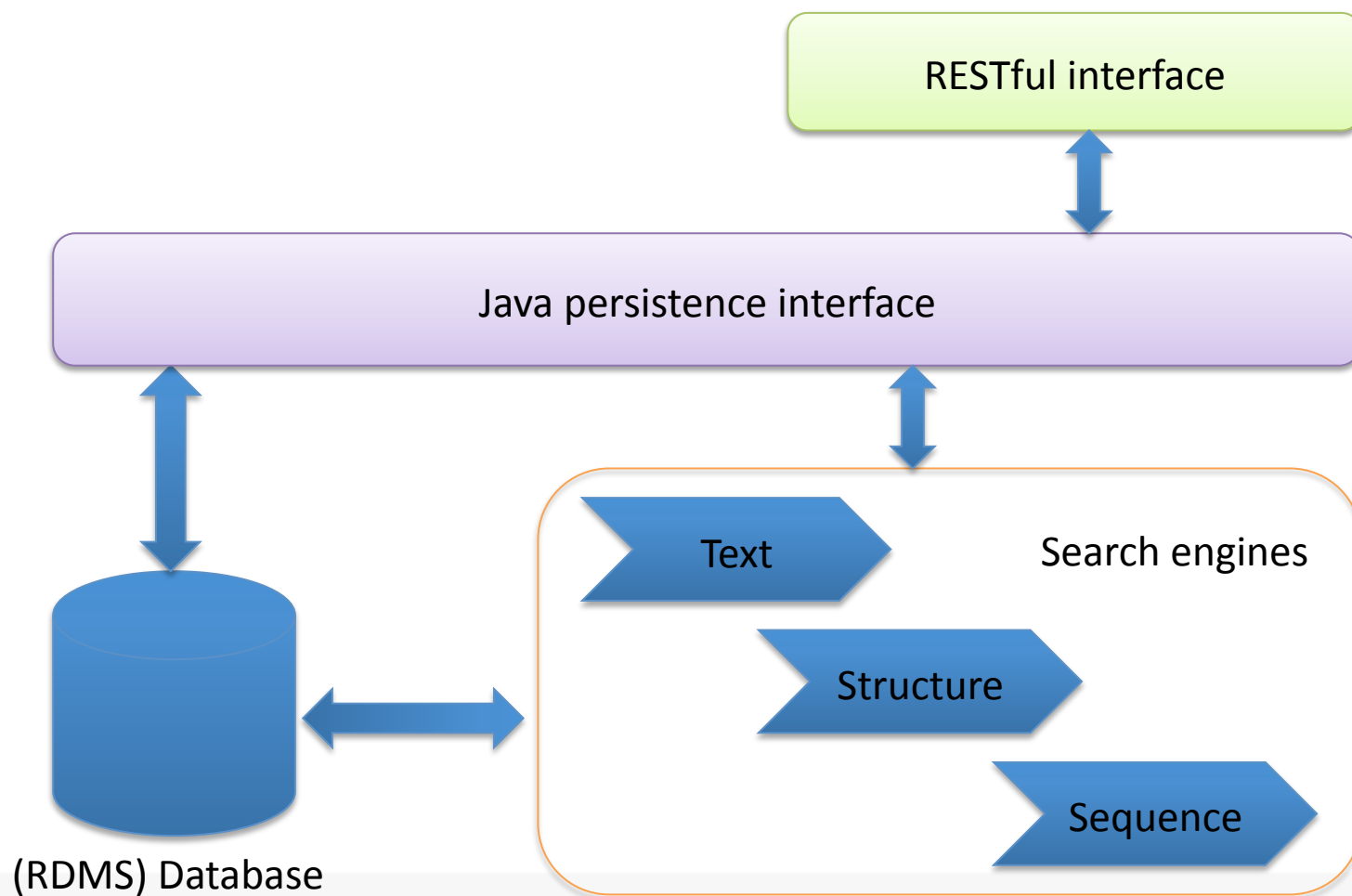
Implementation Roadmap

- Multiple stages
- Stage I (alpha)
 - » Use cases limited to FDA & CBG-MEB
 - » Solidify backend and data models
 - » Chemical substance
- Stage II (beta)
 - » Limited deployment at FDA
 - » Support polymer & protein substances
- Stage III (beta)
 - » Support for remaining substance categories
 - » ISO 11238 compliance
 - » Open deployments to other organizations
- Stage IV (public release)

Architecture Overview

- Client-server
 - » Server is self-hosting when installed on desktop
- Backend database agnostic (e.g., Oracle, MySQL)
- Support two APIs: Native (Java) and RESTful
- Standalone server or deployable within a standard web container (e.g., Glassfish, Tomcat)
- Pluggable chemical toolkit
 - » Default toolkit is a redistributable version of JChem
- Pluggable engines for text, structure, and sequence searching

Architecture Overview (cont'd)



Technology Stack

- Web-based client
 - » Combination of client- and server-side technologies (e.g., ExtJS, JSF)
- Desktop client
 - » Java Swing and other open source libraries
 - » Deploy as either signed webstart or installed image
- Server
 - » JDO as the persistence layer
 - » Lucene text search engine
 - » Custom implementations of structure and sequence search engines
 - » Standalone server based on embedded Jetty or Glassfish and H2 database

Status & Milestones

- Basic user interface framework for web and desktop clients in place
- On-going development on the server backend
 - » Solidify the Java persistence API
 - » Develop initial set of “business rules”
 - » RESTful API
- Anticipate completion of Stage I in June 2013
- If all goes well, Stage II by early 2014

Discussion

- Organic development model
 - » Features evolve through iterations
- Code is hosted at bitbucket.org
 - » Repository is currently private but access is available upon request
 - » Repository goes public in Stage II (beta)
- Demos available for web and desktop clients

Web Client

OpenSRS

Search Register Approve Administration

Enter New Substance

Substance Type

Chemical

Protein

Nucleic Acid

Polymer

Structurally Diverse

Mixture

Entering

Similar (0)

Status

Chemical Structure

New

Lasso

Single

Template

Nitrogen

Center

Ink

No JSDraw License Found

NCC1=CC=C2C(=C1)OC3CCOC3=C2

Stereochemistry

Type

ACHIRAL

Optical Activity

()

Web Client

OpenSRS

Search

Register

Approve

Administration

Search Substance

By Name

By Structure

Advanced

New

Undo

Single

Template

Oxygen

Center

No JSDraw License Found

Oc1ccc2ccccc2c1

☒ Substructure

☐ Superstructure

☐ Similarity

Search

Search Results

CN(C)C(O)COc1ccccc1

Oc1ccc2ccccc2c1

CN(C)C(O)COc1ccccc1

NIH National Center
for Advancing
Translational Sciences

NCATS

Desktop Client

Curation status

Edit trail

Link out to INN document
http://whqlibdoc.who.int/inn/proposed_lists/prop_INN_list77.pdf#page=5

Preferred structure

Registered instances; missing stereocenters annotated

Target information

The screenshot displays the OpenSRS Desktop Client interface. The main window shows the chemical structure of Eplerenone. To the left, a sidebar contains a 'CONTENTS' section with icons for Substances, Proteins, Nucleic Acid, Polymers, and Structurally Diverse. Below this is a 'FILTERS' section with 'Stereo (defined)' and 'Stereo (undefined)' options. The 'COLLECTIONS' section includes 'My favorites' and 'PD2'. The 'TASKS' section is also visible. The main window has a 'Name' field with 'Eplerenone' and a 'Preferred structure' label. Below the structure, there are 'Identifiers' and 'Properties' tabs. The 'Identifiers' tab shows 'Eplerenone', 'Epoxymexrenone', 'INSpra', 'methyl hydrogen 9,11a-epoxy-17a-hydroxy-3-oxopregn-4-ene-7a,21-dicarboxylate', 'Pharmacia brand of eplerenone', 'Cgp-30083', 'Cgp 30083', 'Methyl hydrogen 9,11a-epoxy-17a-hydroxy-3-oxopregn-4-ene-7a,21-dicarboxylate', 'Eplerenonum', and 'Inspra'. The 'Properties' tab shows 'Synonyms', 'Collections', and 'Filters'. At the bottom, there are three chemical structures, with the third one highlighted. A mass spectrum plot is shown in the bottom left corner, with peaks at 240.7, 450.6, and 474.972.3. The x-axis is labeled 'm/z' and the y-axis is labeled 'Norm.'. The plot title is '*MS - 1.835:1.900 min' and the maximum value is 'Max: 18676'.

Desktop Client

Text or structure
searching

OpenSRS [DEVEL] — Substances (16,104 total)

sub:O1C=NC2=CC=CC=C12

CONTENTS

- Substances 16
- Proteins
- Nucleic Acid
- Polymers
- Structurally Dive... 502

FILTERS

- Stereo (defined) 490
- Stereo (undefined)

COLLECTIONS

- My favorites
- PD2

TASKS

Oxadimidine (inn)	Quazolast	Pd-196860	Eclazolast
Flunoxaprofen	Ontazolast	Naftoxatum	Bifeprunox
Chlorzoxazone	Zoxazolaminum	NCGC00016280-07	NCGC00015238-13

16 substances